Technologies for the ProLiant ML570 G3 and ProLiant DL580 G3 servers

technology brief

Abstract

With the third-generation of ProLiant 500-series servers —the ProLiant ML570 G3 and the ProLiant DL580 G3 servers — IT managers have additional choices for servers that can transition easily from 32-bit to 64-bit x86 processing. The architecture for the ProLiant ML570 G3 and the DL580 G3 uses the Intel® 64-bit Xeon™ processor MP designed to operate in either 32-bit or 64-bit mode, depending on the application and operating system (OS) utilized. This architecture brings enhanced performance not only through the 64-bit architecture but also through the use of fast DDR-2 memory, large memory footprints, and dual front-side buses. To complement this high-performing architecture, the Intel E8500 chipset adds high-availability memory technologies such as Online Spare, Hot Plug Mirrored memory, and Hot Plug RAID memory. This generation of servers brings Hot Plug RAID technology — providing the highest level of availability, at a lower cost than mirroring — into the 4-way multiprocessing, mid-tier enterprise solution space.

The ProLiant ML570 G3 and the DL580 G3 servers have been completely redesigned to provide increased serviceability and reliability in their mechanical designs. The chassis designs have minimal cabling, fool-proof locking mechanisms to avoid mishaps, simple rack rail systems, and are virtually tool-less.

Introduction

The ProLiant ML570 G3 and the ProLiant DL580 G3 servers bring a new level of high-performance and high-availability technologies to 4-way, industry-standard servers. Many of these technologies were previously available only in 8-way x86 servers or servers using other processor architectures.

This brief addresses the following key technologies within the ProLiant ML570 G3 and DL580 G3 platforms:

- The move to 64-bit processor operations
- Faster, higher-performing memory
- High-availability technologies including Advanced Memory Protection technologies and larger memory footprints
- Updated I/O technologies
- Highly leveraged mechanical designs to ensure serviceability

This brief discusses only certain key technologies of the ProLiant 500-series servers. For complete specifications of each server, see the HP website at www.hp.com/go/proliant.

Processor architecture

In 2004, HP introduced the first 4-way ProLiant server to provide 64-bit x86 capabilities —the ProLiant DL585 using AMD Opteron processors. The ProLiant ML570 G3 and the ProLiant DL580 G3, announced in March 2005, extend the 64-bit portfolio with their use of the Intel® 64-bit Xeon™ Processor MP with up to 8 MB L3 cache. Its use of Intel Extended Memory 64-bit Technology (EM64T) enables IT organizations to deploy common platforms for both 32-bit and 64-bit computing, and move to 64-bit computing gradually as it benefits their businesses.

Defining 64-bit architecture

A 64-bit architecture has a much larger amount of directly addressable (flat) memory space than a 32-bit processor. The use of EM64T allows the OS to access a flat memory address space greater than 4 GB without enabling Physical Address Extensions (PAE) and incurring the overhead of PAE.

This can result in performance advantages for the 64-bit architectures because of their ability to use large amounts of memory, such as with intensive floating-point calculations used in scientific and engineering modeling programs.

For additional information about 64-bit extensions and architecture, see the technology brief[1] titled "Characterizing x86 processors for industry-standard servers: AMD Opteron and Intel Xeon"

Xeon processor MP

The 64-bit Xeon processor MP comes in two different versions: a version with a 1 MB L2 cache; and a version with up to 8 MB of L3 cache in addition to the 1 MB L2 cache. Both are built using 90 nm process technology and use a 166-MHz front side bus which is quad-pumped to 667 MHz, providing up to 5.3 GB/s of data transfer rates. The processors support IA-32 and the EMT64 instruction set for running 64-bit applications and operating systems.

As of this writing, the ProLiant ML570 G3 and DL580 G3 platforms support the following processors:

- Xeon MP 3.3 GHz/8 MB L3/1 MB L2
- Xeon MP 3.0 GHz/8 MB L3/1 MB L2
- Xeon MP 2.83 GHz /4MB L3/1 MB L2
- Xeon MP 3.66 GHz/1 MB L2
- Xeon MP 3.16 GHz/1 MB L2

The 64-bit Xeon processor MP uses the NetBurst architecture with Hyper-Threading technology, Hyper-Pipelined technology, and a 12K Execution Trace Cache. It includes support for Enhanced Intel Speed-Step Technology and Intel Execute Disable Bit technology.

As server and rack densities have increased, power and heat management are becoming increasingly important. In response, Intel developed Enhanced Intel Speed-Step Technology, which exposes power state registers in the processor. With the appropriate ROM or OS interface, these registers can be used to switch the processor between different power states, changing the processor's operating frequency and voltage. This, in turn, lowers the power usage and heat production of the processor. Demand-based switching is the OS implementation of power management using the Enhanced Speed-Step technology, and is supported by some new operating systems including Microsoft Windows Server 2003 SP1, Red Hat Enterprise Linux 4 Update 1, and SUSE Linux Enterprise Server 9 SP1.

HP Power Regulator for ProLiant[2] is an OS–independent power management feature of HP ProLiant servers that uses Enhanced Speed-Step technology. HP Power Regulator supports both dynamic and static modes. With HP Static Low Power Mode, the processors are configured to run continuously in a lower power state. This is useful for customers with power-constrained data centers who require a guaranteed maximum power usage for each server. For servers that operate in moderately or minimally loaded environments, there will be little, if any, performance degradation. HP Dynamic Power Savings mode lowers overall power usage of the server without affecting system performance. When this feature is enabled,[3] the System ROM will dynamically modify each processor's frequency and voltage based on the processor workload. The processor operates in a high power state only when needed, thus reducing the overall system power usage.

Intel first released the Execute Disable Bit functionality with the Itanium processor family. The technology allows the processor to classify areas of memory which cannot execute application code. When combined with OS support, this helps to prevent certain classes of malicious buffer overflow attacks. As of this writing, Intel Execute Disable bit is supported by Microsoft Windows Server

[1] Available on the ISS Technology Papers website at http://h18004.www1.hp.com/products/servers/technology/whitepapers/

[2] For additional information about Power Regulator for ProLiant, see http://h18000.www1.hp.com/products/servers/management/ilo/power-regulator.html

[3] The ProLiant ML570 G3 and ML580 G3 system ROMs are expected to support HP Power Regulator mid-year 2005.

2003 SP1, Microsoft Windows XP SP2, SuSe Linux Enterprise Server 9.2, or Red Hat Enterprise Linux 3 Update 3.

For additional information about these processors, see the Intel website or the technology brief titled "The Intel processor roadmap for industry-standard servers."[4]

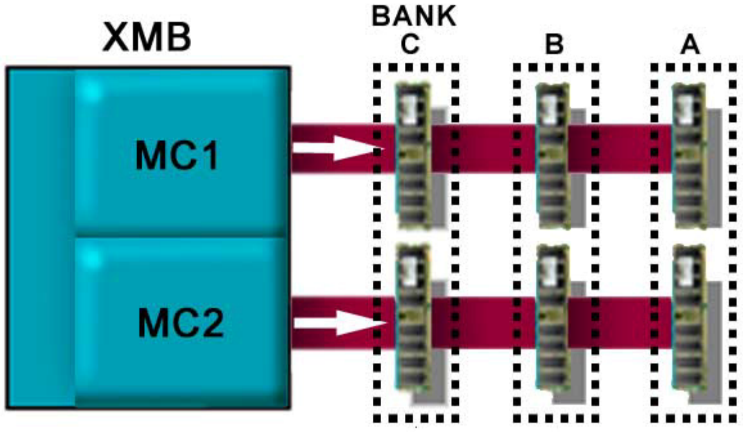
Memory architecture

The ProLiant ML570 G3 and the ProLiant DL580 G3 both use the Intel E8500 chipset architecture, resulting in the same core technologies for the memory subsystem. However, they differ in their I/O implementation to meet diverse customer requirements.

Background: memory banks and ranks

The term memory bank has been used to refer to more than one concept. For this paper, a memory bank refers to a pair of DIMMs that are located in the same order in two parallel memory channels (Figure 1). The DIMMs may be single-rank or dual-rank, which affects memory capacity and how memory is interleaved for performance.

Figure 1. A memory bank is a pair of DIMMs. The ProLiant ML570 G3 has three memory banks; the ProLiant DL580 G3 has two.



A single-rank DIMM is a DIMM in which all of the memory chips contribute to a single data set of 64 bits (plus the ECC bits) and are activated by the same chip-select signals.

To increase memory density, memory suppliers are producing dual-rank DIMMs. Typically, a dual-rank DIMM is made by stacking a second set of memory chips directly on top of the first set of memory chips. A dual-rank DIMM produces a second data set of 64 bits (plus ECC bits) and requires two chip-selects with different signals to differentiate between the two sets of memory chips. Although physically taking up the space of a single DIMM, a dual-rank DIMM acts as if it were two separate DIMMs, and is considered two electrical loads by the chipset.

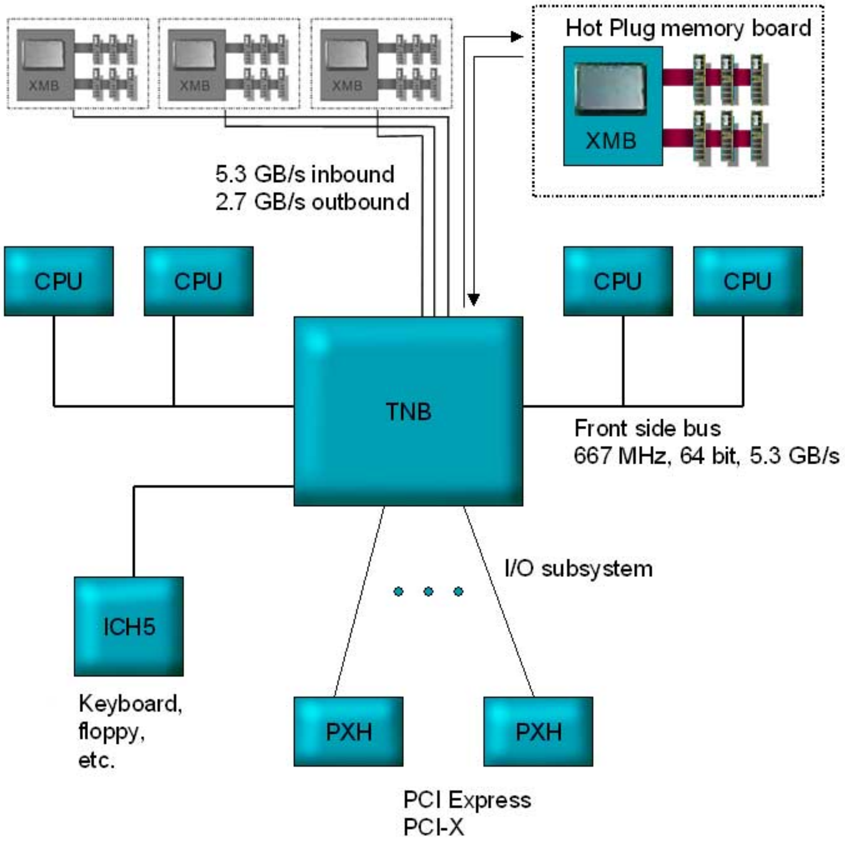
[4] This technology brief available on the ISS Technology Papers website at http://h18004.www1.hp.com/products/servers/technology/whitepapers/adv-technology.html

Intel E8500 chipset

The E8500 chipset has a high-availability memory subsystem that consists of the North Bridge (TNB) and the XMB memory controller.

The E8500 chipset is designed to support the upcoming dual-core versions of Intel processors, and has the important feature of using two separate front side buses to connect to the processors (Figure 2). Each north bridge can connect to up to four memory boards, and each memory board includes an XMB memory controller chip. The north bridge connects to each XMB memory controller using a high-speed serial interconnect (the IMI bus) that allows 5.3 GB/s inbound (for read data signals from the XMB) and 2.7 GB/s outbound (for write data signals to the XMB). This inbound speed matches the throughput of the front-side buses, providing a good balance between the processor and memory subsystem. The north bridge uses an in-order (FIFO) queue to maintain coherency across the dual front-side buses while processing read/write requests.

Figure 2. Xeon MP architecture used in the ProLiant ML570 G3 and DL580 G3 platforms

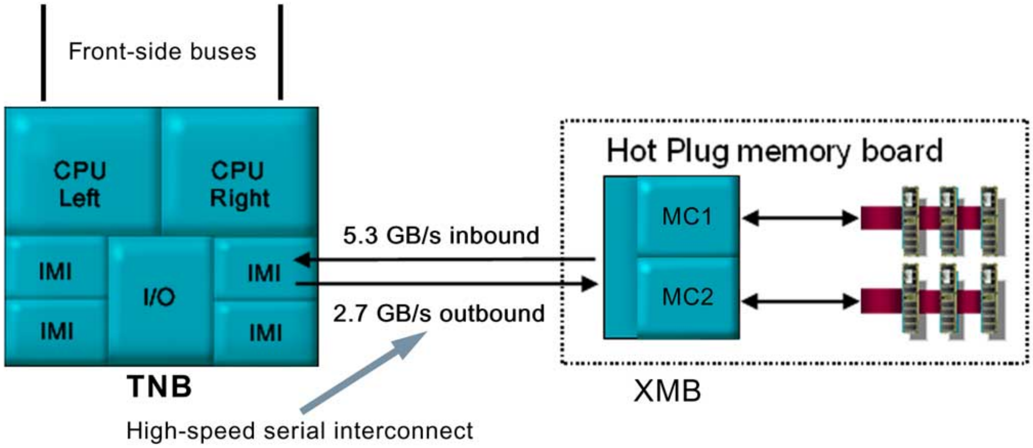


Each XMB memory controller chip supports two channels of DDR-2 memory. The DDR-2 memory on each channel operates in lockstep at 400 MHz. The ProLiant ML570 G3 supports 6 DIMMs per memory board (three per channel), and the ProLiant DL580 supports 4 DIMMs per memory board (two per channel), due to physical constraints of the 4U system. For both servers, the maximum memory supported is 64 GB with 4-GB DIMMs, as described in the section "Maximum memory configurations."

Partitioning for electrical isolation

One of the features of a well-designed chipset is the degree to which the silicon is partitioned to allow different signal areas to be electrically isolated. The E8500 chipset is partitioned so that the front-side bus interconnects to a partitioned area for the "left" CPU, the "right" CPU, and each memory board (Figure 3). The XMB is similarly partitioned so that each internal memory controller is isolated electrically from the other to avoid power noise and crosstalk issues. Avoiding crosstalk and other noise is increasingly important as bus speeds increase and bus signals become more susceptible to slight differences in voltages.

Figure 3. Example of how the TNB and XMB chips are partitioned to reduce power noise and crosstalk issues.



Maximum memory configurations

Each XMB memory controller supports eight electrical loads. A single-rank DIMM is considered one electrical load; a dual-rank DIMM is two electrical loads. Therefore, the ProLiant ML570 G3 supports the following maximum DIMM configurations per memory board:

- Six single-rank DIMMs (three per memory channel)
- Four dual-rank DIMMs (two per memory channel)
- Two dual-rank DIMMS and four singe-rank DIMMs

When 4-GB, dual-rank DIMMs are available, a customer can use four dual-rank DIMMs per memory board to provide the maximum memory of 64 GB for the ProLiant ML570 G3.

The ProLiant DL580 G3 also supports a maximum of 64 GB of memory using four, dual-rank, 4-GB DIMMs per memory board. The system can support a maximum of four DIMMs per memory board, using either single-rank DIMMs, dual-rank DIMMs, or a combination of the two.

In either system, DIMMs must be installed in pairs on the memory board. Each DIMM pair must be identical, with the same capacity, technology, and density. Refer to the server's user guide for valid memory configurations when combining single and dual-rank DIMMs.

High-performance memory

Processor performance has kept pace fairly consistently with Moore's law of doubling performance every two years. On the other hand, memory bandwidth doubles roughly every three years. To keep pace, designers are challenged to make memory subsystems that are faster. The ProLiant ML570 G3 and the ProLiant DL580 G3 use DDR2-400 memory and interleaving to improve memory performance and decrease this performance gap.

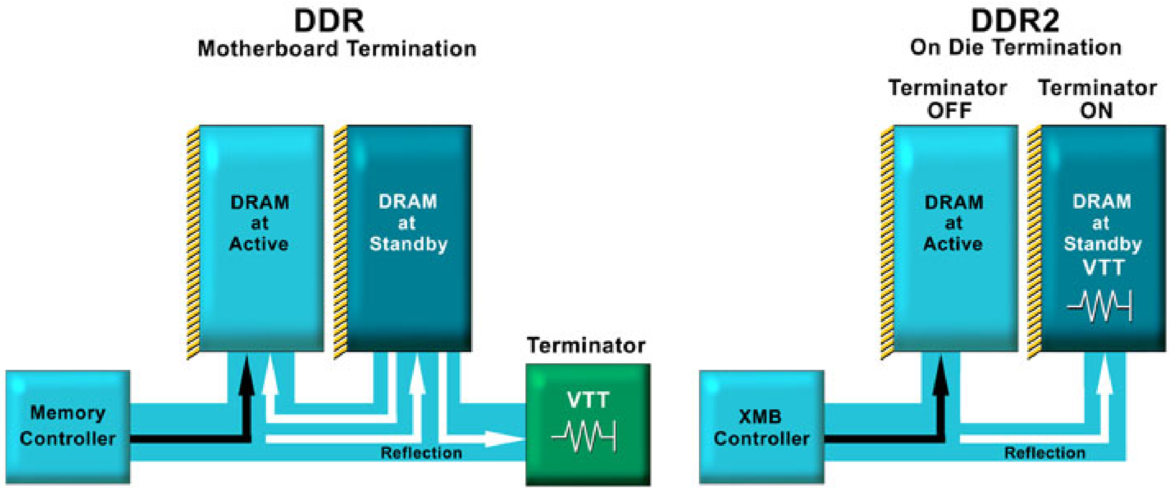
DDR-2 memory

DDR-2 SDRAM is the second generation of DDR SDRAM. In contrast to the first generation, DDR-2 memory operates at an even lower voltage (1.8V) to further reduce power consumption; uses higher clock frequencies to increase data transfer rates, and uses on-die termination control to improve signal quality. At 200 MHz (double-clocked to an effective 400 MHz), DDR-2 increases memory bandwidth to 3.2 GB/s.

In DDR memory, an external termination resistor is added to the system board to improve signal quality and reduce noise for the memory signals. This reduces the likelihood of a signal reflecting, or bouncing back, toward the driving source. DDR-2 memory, on the other hand, moves this resistor into the silicon of the memory itself. This reduces signal reflection and therefore improves signal quality (Figure 4).

Refer to the technology brief titled "Memory technology evolution: an overview of system memory technologies" for additional information about DDR-2 memory technology.[5]

Figure 4. On-die termination reduces the amount of signal reflection to improve signal quality.



Memory interleaving

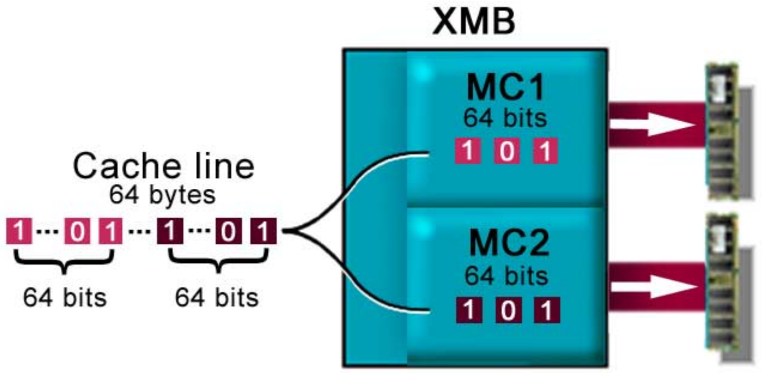
To reduce latencies and improve performance, there are three different types of memory interleaving within the ProLiant ML570 G3 and DL580 G3 servers: two-way (dual channel) interleaving, interleaving within the XMB memory controller, and interleaving across multiple XMB memory controllers. To simplify the descriptions, the following sections describe how interleaving works when using single-rank DIMMs. The same concepts apply for dual-rank DIMMs.

[5] Available on the ISS Technology papers website at http://h18004.www1.hp.com/products/servers/technology/whitepapers/adv-technology.html.

Two-way interleaving

Like previous ProLiant servers, the ML570 G3 and the DL580 G3 servers use two-way, or dual-channel, interleaving. Two-way interleaving works by dividing memory into 64-bit blocks that can be accessed two at a time through the two memory channels in an XMB controller (Figure 5). This results in twice the amount of data obtained in a single memory access and reduces the required number of memory accesses. Because the data is split into the two separate memory channels and accessed simultaneously, DIMMs must be installed in pairs in the ProLiant ML570 G3 and DL580 G3 servers, and the pairs must contain identical DIMMs.

Figure 5. Interleaving between the two channels of memory allows 64-bits to go to memory controller (channel) 1, then the next 64 bits to go to memory controller 2, and so on.

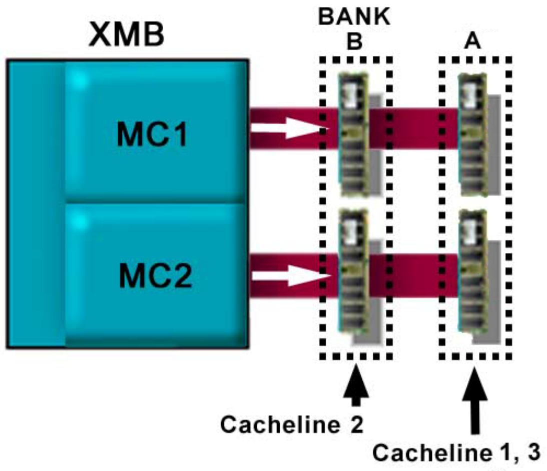


XMB rank interleaving

Rank interleaving within the XMB groups several ranks of memory together so that cache lines are sequentially read or written across the entire group. For example, suppose that bank A and bank B are interleaved together. (In this example, all contain single-rank DIMMs, so bank A is equivalent to a rank.) The first requested cache line would come from bank A DIMMs, then the next cache line from bank B DIMMs, the next from bank A DIMMs, and so on (Figure 6). XMB interleaving is done on 2, 4, or 8 ranks at a time.

XMB rank interleaving reduces latencies by allowing multiple memory pages to be open at the same time, rather than waiting for several cache lines to be read from a single bank.

Figure 6. Rank interleaving splits cache lines across a group of memory ranks.



Interleaving across XMB memory controllers

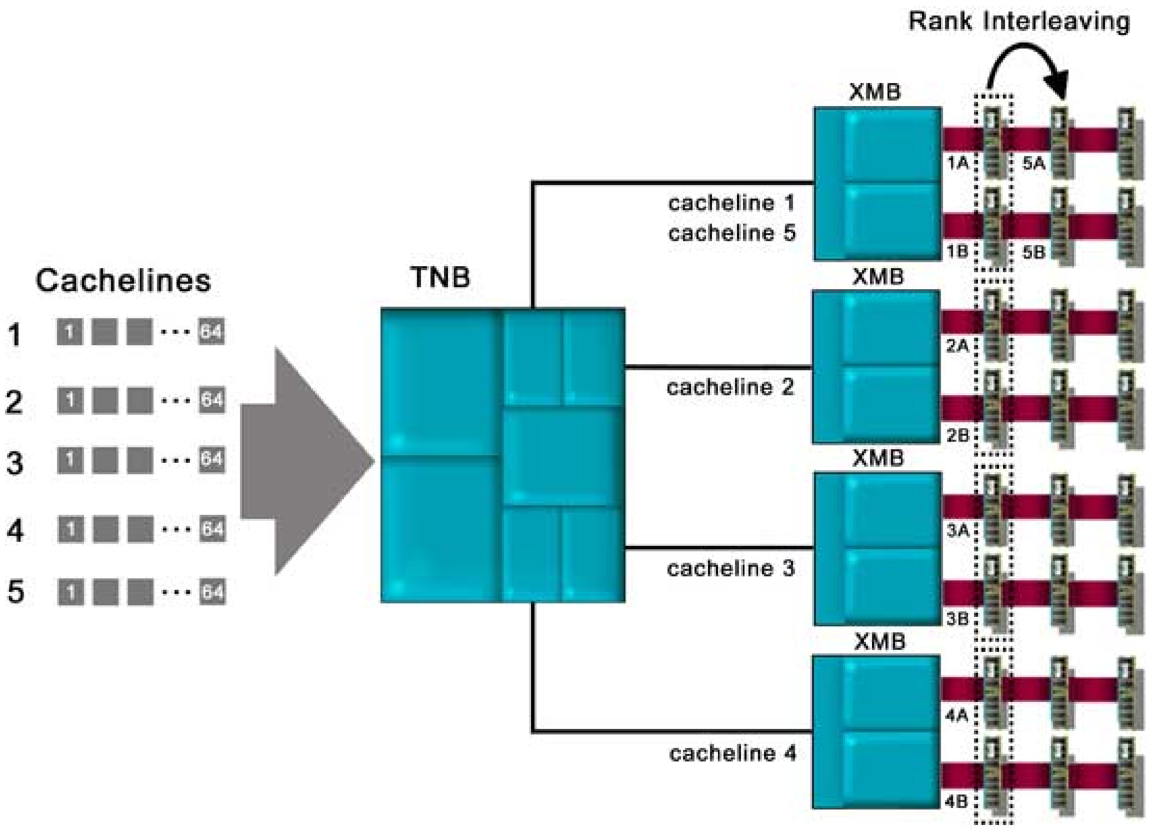
Because there are up to four XMB memory controllers (one per memory board), the north bridge device also splits requested cache lines among all memory controllers. This type of interleaving is automatically enabled when using more than one memory board (if hot add is disabled).

For example, if an application requires 40 cache lines of data and the server contains four memory boards, the north bridge will split those 40 cache lines across the four memory boards in a sequential fashion: cache line 1 comes from memory board 1, cache line 2 comes from memory board 2, and so on (Figure 7). Assuming that rank interleaving is also enabled, the memory controller will read the fifth cache line from the second rank of DIMMs on memory board 1. Distributing the memory reads and writes across multiple controllers reduces memory access times for individual controllers and reduces the likelihood of processors waiting on data.

The north bridge has no mechanism for updating the interleaving scheme on-the-fly if new memory boards are added. Therefore, if hot-add memory is enabled, interleaving cannot be done across memory controllers and must be disabled.[6]

[6] The one exception is if all four memory boards are installed and hot-add is enabled: because the system is already fully populated with memory boards, the north bridge controller can interleave memory across the XMB memory controllers.

Figure 7. With interleaving across XMB memory controllers, the north bridge interleaves the cache lines among the memory board subsystems (moving down vertically in the schematic). Rank interleaving would then interleave cache lines horizontally, as depicted with cache line 5.



Errors in memory

The move to a 64-bit architecture naturally leads to servers that support more memory capacity to fully utilize the capabilities of the 64-bit architecture. The continued reduction in cost of high-capacity memory modules also leads customers to install more memory as a way of achieving high performance. However, as memory capacity grows, it becomes statistically more likely that memory errors will occur: both hard and soft errors.

Hard and soft errors

A hard memory error is characterized by the fact that it is repeatable and indicates a physical problem such as a memory defect or a broken connection on the DIMM. The data in a DIMM that contains a hard error may be corrected using standard or advanced ECC (depending on the number of bits in error). However, the error itself cannot be fixed and every time that memory location is read, another error occurs.

Most errors that occur in the memory subsystem are soft errors. A soft error is a randomly occurring event caused by external influences such as high-energy alpha particles and cosmic radiation that has enough energy to penetrate the earth's surface. When such particles collide with a memory storage device, it may disturb the state of the data bit(s). According to the JEDEC standard[7], soft error rates

[7] The JEDEC Solid State Technology Assocation is the prominent developer of standards in the solid state electronics industry. The JEDEC standard JESD89, "Measurement and Reporting of Alpha Particles and terrestrial Cosmic-Ray Soft Errors Induced in Semiconductor Devices," is available. at www.jedec.org

are affected by the increased density of memory devices, system voltage and timing margins, memory system operating frequencies, radioisotropic impurities in packaging and circuit board materials, and even magnetic variations due to altitude/location around the earth.

Soft errors can be corrected using standard or advanced ECC. Because a soft error is not caused by a problem with the DIMM, once the data is corrected, the same error will not recur in the same component.

Correctable and uncorrectable errors

Errors can be categorized as either correctable or uncorrectable. In the ProLiant ML570 G3 and the DL580 G3 servers, the memory controller calculates check bits every time it writes to memory. When memory is read, it re-calculates those check bits from the data stored in the DRAM devices and compares the re-calculated check bits to the stored check bits. If the two sets of check bits are different, the error can be corrected if it is a:

- Single-bit error in a DRAM device (correctable by standard ECC)
- Multi-bit error in a DRAM device (correctable by advanced ECC)

If multi-bit failures occur in different DRAM devices, they are not correctable. The uncorrectable error will return bad data unless the customer has enabled Advanced Memory Protection techniques such as Hot Plug RAID or Hot Plug Mirrored Memory.

Standard ECC

To significantly reduce the probability of fatal memory failures, HP was the first company to introduce ECC memory in industry-standard servers in 1993. ECC memory is now standard in all HP ProLiant servers and most other industry-standard servers. Standard ECC detects both single-bit and double-bit errors, and it corrects single-bit errors.

Advanced ECC (single device data correction)

To improve memory protection, HP introduced Advanced ECC technology[8] in 1996. HP and most other server manufacturers continue to use this solution in industry-standard products. The advanced ECC algorithm that Intel uses in the XMB controllers is referred to as single device data correction (SDDC). The eight-bit (x8) implementation of SDDC can detect and correct multi-bit failures in a four-bit (x4) or x8 DRAM device,[9] which makes it able to recover from a x4 or x8 DRAM component failure. It can also detect errors in two x4 DRAM components.

Advanced ECC is the only memory protection technique for the ProLiant ML570 G3 and the ProLiant DL580 G3 servers that supports hot-add. Hot-add refers to adding memory boards to the system while it is running, which make additional memory resources available to the OS. It must be enabled in the ROM-Based Setup Utility (RBSU) and must be supported by the OS. Advanced ECC with hot-add enabled allows the amount of memory available to the OS to be increased without rebooting the system.

Demand scrubbing

After the chipset detects a correctable memory error, it recalculates the correct data using ECC (or advanced ECC) check bits and sends this correct data back to the processor. For soft errors, the invalid data is still present in the DRAM unless that memory error is scrubbed, or the good data is written back to the DRAM. The ProLiant ML570 G3 and DL580 G3 servers support a memory scrubbing technique called demand scrubbing.

[8] U.S. Patent assigned to HP. D.G. Abdoo and J.D. Cabello, "Error Correction System for N Bits Using Error Correcting Code Designed for Fewer than N Bits." U.S. Patent 5,490,155 (Feb. 6, 1996).

[9] Server memory DIMMs use DRAM chips which hold either 4 or 8 bits, known as x4 or x8 devices.

Demand scrubbing allows the chipset to write back good data on a memory read if a correctable memory error is detected. If future reads occurred at that same memory location without the data being scrubbed, the chipset would detect another correctable error, which may result in the system marking the DIMM as degraded. For soft errors, demand scrubbing will prevent all subsequent correctable errors after the first error is encountered. Demand scrubbing also reduces the likelihood of another soft error occurring, resulting in a multi-bit error. Multi-bit errors cause a system failure if Hot Plug Mirrored Memory or Hot Plug RAID is not enabled on the system.

The ProLiant ML570 G3 and ProLiant DL580 G3 servers support demand scrubbing in system ROMs dated after Feb 28, 2005.

High-availability memory technologies

The ProLiant ML570 G3 and DL580 G3 servers offer three levels of Advanced Memory Protection that provide increased fault tolerance for applications requiring higher levels of availability: Online Spare; Hot Plug Mirrored Memory; and Hot Plug RAID.

Hot-plug definitions

As already mentioned, advanced ECC supports hot-add of memory boards so that the amount of memory available to the OS is increased while the server is running.

Hot-replace, on the other hand, allows a memory board to be removed, the failed or degraded DIMMs to be replaced, and the memory board to be re-installed, all while the server is running. It is available without any OS support and can be used with either mirroring or RAID techniques.

Online Spare

With Online Spare mode, when a server DIMM exceeds a threshold rate of correctable memory errors, that rank of memory within the DIMM that has exceeded the threshold is taken offline and the XMB memory controller copies the data to a replacement rank (the Online Spare). Because a DIMM that has a high rate of correctable memory errors is at an increased risk of having an uncorrectable memory error, Online Spare allows the user to remove these higher-risk DIMMs from the memory map. Using Online Spare reduces the chance of an uncorrectable error bringing down the system; however, it does not fully protect the system against uncorrectable memory errors.

When a system uses Online Spare memory, the Online Spare rank must be at least as large as all other memory ranks on the memory board. In Online Spare mode, one rank of memory per memory board is reserved for the spare rank and is not available to the OS. For a memory board containing varying sizes of DIMMs, the system chooses the largest rank on the memory board as the Online Spare.

Online Spare works independently for each memory board. In other words, each board can copy data to its Online Spare rank independent of what is happening with any other memory board. It is supported with any number of memory boards installed.

Online Spare memory does not support any hot-plug operations. While the server must still be powered down to replace a bad memory module, the server can continue to operate until a scheduled shutdown.

Hot Plug Mirrored Memory

Mirrored memory mode is a fault-tolerant memory option that provides a higher level of availability than Online Spare memory. Mirrored memory allows the server to keep two copies of all memory data on separate memory boards. This allows the system to be protected against uncorrectable memory errors. If an uncorrectable error is encountered, then the server automatically retrieves correct data from the memory board that does not contain errors.

The ProLiant ML570 G3 and the DL580 G3 support mirroring of two or four memory boards, allowing the server to be completely protected from memory failures. However, the customer effectively uses half of the installed memory capacity (for example, if 4 GB is installed, only 2GB is available to the OS and applications).

The mirroring functionality requires no OS support but does require each installed memory board to have the same total amount of memory. The mirroring function supports hot-replace without any OS support. For this reason, mirrored memory mode is beneficial to businesses that cannot afford downtime and cannot risk waiting until scheduled downtime to replace degraded memory modules.

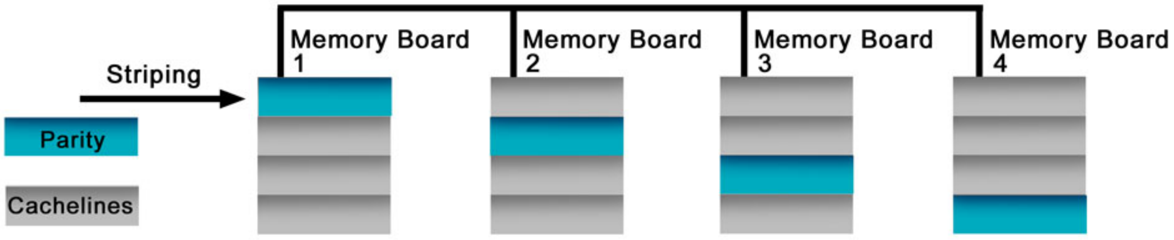
Hot Plug RAID

HP introduced Hot Plug RAID memory in its 8-way ProLiant servers using the F8 architecture. Hot Plug RAID memory protects the server against uncorrectable memory errors that would otherwise result in a server failure. If Hot Plug RAID memory is enabled, then the north bridge uses an exclusive-OR engine to generate a parity check line for every three cache lines. The north bridge interleaves the cache lines and the parity check line across all four memory boards (Figure 8). If an uncorrectable memory error is encountered, the server can re-create the proper data using the parity information and the information from the other memory boards that contain no failures.

Therefore, to use Hot Plug RAID memory, administrators must install all four memory boards in the ProLiant ML570 G3 and the DL580 G3 servers. Because the cache lines are striped across the memory boards, all four boards must have the same total amount of memory.

As with Hot Plug Mirrored Memory, Hot Plug RAID memory allows failed or degraded DIMMs to be replaced while the server is running without requiring server downtime. The memory board with the failed DIMMs can be removed, failed DIMMs replaced, and the board re-inserted into the server without any interruption to the OS. Furthermore, using Hot Plug RAID provides the most cost-effect means of protecting against uncorrectable errors, because only 25 percent of the memory is allocated to parity information.

Figure 8. Hot Plug RAID memory for the ProLiant ML570 G3 and DL580 G3 stripes three cache lines and the associated parity information across the memory boards.



Comparing Advanced Memory Protection technologies

Customers have several options to consider when evaluating the different memory protection options in the ProLiant ML570 G3 and the DL580 G3 servers.

Generally, as the level of memory availability (redundancy) increases, the amount of installed memory available for use by the OS decreases. Advanced ECC provides the most available memory, as all installed memory is available to the OS and applications. Advanced ECC mode protects against correctable memory errors. However, advanced ECC mode does not protect against uncorrectable errors and does not provide any capability of replacing failed or degraded DIMMs without shutting down the server.

Online Spare memory reduces the likelihood of uncorrectable memory errors but does not protect against uncorrectable memory errors. Like advanced ECC, Online Spare provides no capability of replacing failed or degraded DIMMs without shutting down the server. The amount of memory reserved for the Online Spare rank will vary from one system configuration to the next. The exact percentage of memory available to the OS when using Online Spare mode depends on the number and size of DIMMs populated per memory board.

Hot Plug Mirrored Memory provides protection against both correctable and uncorrectable memory errors. Hot Plug Mirrored Memory also allows replacing failed or degraded DIMMs while the system is operating. However, this mode uses half of the installed memory for redundancy.

Hot Plug RAID memory often provides the most economical and effective memory protection. Like Hot Plug Mirrored Memory, it protects against correctable and uncorrectable memory errors, but does so while allowing 75 percent of the installed memory to be available to the OS. Hot Plug RAID memory also allows replacing failed or degraded DIMMs while the system is operating.

Table 1 summarizes the choices between levels of memory protection.

Table 1. Tradeoffs between hot add, hot replace, and amount of memory available for system use.

Memory Option	Hot-add support	Hot-replace support	Memory utilization
Advanced ECC	Yes	No	All
Online Spare	No	No	Varies
Mirroring	No	Yes	50%
RAID	No	Yes	75%

Ensuring reliability within large-footprint memory designs

High-availability memory technologies are one way to assure the customer that the system will be reliable as memory footprints grow larger. Another way to ensure reliability is to carefully choose memory suppliers and components.

One of the challenges with increasing amounts of memory is the variation among memory suppliers. Variations can occur at the design level as well as the production level. For example, although many of the design parameters for a DIMM design are specified by JEDEC, the design of the I/O buffer which drives and receives the signal is not specified. Therefore, each vendor designs their own buffers, which determines the signal strength that is sent to or from memory. In addition, over the life of a particular DIMM technology (for example, 512-MB DIMMs using DDR technology at 133 MHz), the DIMM suppliers may change internal silicon designs to increase yield or fix minor bugs. Any of these variations can cause potential problems in the platform if the variations are excessive.

HP employs multiple tactics to ensure quality memory components. First, HP uses industry-standard components from top-level suppliers and makes sure that these components meet standard specifications. In addition, HP works closely with chipset and component designers to ensure that normal variations among components do not cause reflections or noise that can alter signal integrity. HP engineers use utilities developed in-house to evaluate chipset and system noise margins before a product is released to production. Finally, HP uses several levels of diagnostics during the server manufacturing process: an in-circuit test to verify that components are placed correctly; a functional board test to determine that the motherboard is functioning properly and can boot the operating system; and a full system diagnostic to ensure that all the components in the system function properly.

After the server is shipped to the customer, HP provides Pre-Failure Warranty support standard on all ProLiant servers. HP Systems Insight Manager notifies the administrator when a critical component may fail. During the warranty period, the Pre-Failure Warranty covers the replacement of DIMMs used in a server's main memory when the predefined thresholds for correctable errors have been exceeded. The Pre-Failure Warranty provides for the components to be replaced free of charge under the warranty. With the Pre-Failure Warranty, system administrators can proactively schedule downtime for maintenance and not interrupt critical business operations that rely on these enterprise servers.

Architecture trade-offs with a large memory footprint

The three most recent ProLiant 500-series servers — the ProLiant ML570 G3, ProLiant DL580 G3, and the ProLiant DL585 — all have large memory footprints and high performance architectures (Table 2). Because they use fundamentally different architectures, however, there are tradeoffs that customers should understand when evaluating the three servers.

Table 2. Memory and architecture comparison of ProLiant 500-series servers.

Server	Processor	Architecture	Max. memory support	Memory protection
ProLiant ML570 G3	Intel 64-bit Xeon processor MP	Intel E8500 chipset; dual 667-MHz front side bus; DDR-2 400 memory	48 GB with 2-GB DIMMs; 64 GB with dual-rank, 4-GB DIMMs	Advanced ECC; Online Spare; Mirroring; Hot Plug RAID
ProLiant DL 580 G3	Intel 64-bit Xeon processor MP	Intel E8500 chipset; dual 667-MHz front side bus; DDR-2 400 memory	32 GB with 2-GB DIMMs; 64 GB with dual-rank, 4-GB DIMMs	Advanced ECC; Online Spare; Mirroring; Hot Plug RAID
ProLiant DL585	AMD Opteron 800-series processor	AMD 8000 chipset; HyperTransport 800-MHz or 1-GHz links; DDR1 266, 333, and 400 memory	64 GB with 2-GB DIMMs	Advanced ECC

Both the ProLiant ML570 and the DL580 G3 provide the performance of the latest Intel processor with dual front-side buses and 64-bit extensions. In addition, these servers offer advanced memory protection — important, for example, with customers in a 24x7 business that cannot afford any uncorrectable memory errors that might bring down applications or must have their servers operating at all times without any downtime for replacing or upgrading memory.

The ProLiant DL585 has proven its industry-leading performance since its announcement, and with the announcement of dual-core Opteron processors, will provide even greater performance.[10] The high level of performance is due to its architecture using memory controllers integrated into the processor die, fast HyperTransport links, and the ability to use 64 GB of memory. Like the ProLiant ML570 G3 and DL580 G3 servers, the ProLiant DL585 is capable of using 64-bit applications and operating systems as soon as they are available. The tradeoff for this very fast and scalable performance is that the ProLiant DL585 provides memory protection with advanced ECC only. Advanced ECC memory, however, should not be dismissed as insufficient. It is the standard memory protection technique within the x86 server market, and provides ample protection for the majority of customers.

[10] See the HP website at http://h18004.www1.hp.com/products/servers/benchmarks/ for benchmark performance information.

Updated I/O technologies

Two capabilities are being used in the ProLiant ML570 and DL580 G3 servers which were not present in previous generations of the 500-series servers: PCI Express (available on both servers) and the use of the HP Smart Array 6i storage controller on the ProLiant DL580 G3.

The ProLiant ML570 G3 and DL580 G3 support PCI Express and PCI-X technology to give customers the option to use whichever I/O cards fit their purposes. The ProLiant ML570 G3 includes ten I/O slots: four PCI Express x4, two hot plug 64-bit/133 MHz PCI-X, and four 64-bit/100 MHz PCI-X.

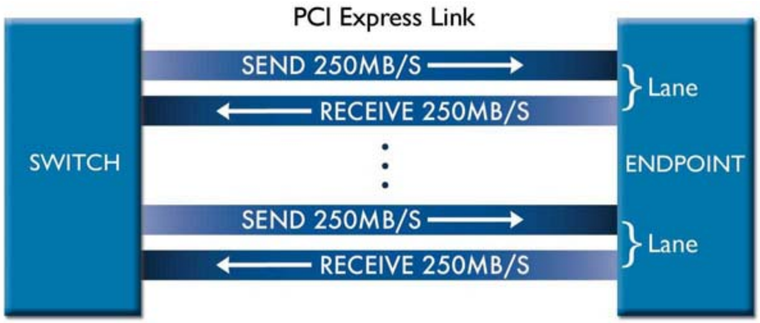
The ProLiant DL580 G3 includes PCI Express slots as mezzanine options, with either two x4 PCI Express slots or a single x8 PCI Express slot. Or, customers can use the mezzanine board that provides two hot-plug PCI-X 64-bit/133 MHz slots. The five standard PCI-X slots in the DL580 G3 include three PCI-X 64-bit/133 MHz and two PCI-X 64-bit/100 MHz.

PCI Express

PCI Express provides point-to-point connections between devices. It sends data serially, one bit after another, over each link rather than sending the data in parallel, one bit beside the other, as in PCI-X. Serial interfaces allow for transmitting data at higher rates as opposed to parallel interfaces.

A PCI Express serial link consists of one or more dual-simplex lanes. Each lane contains a send pair and a receive pair to transmit data at the signaling rate in both directions simultaneously. The serial communication requires an 8b/10b encoder to convert parallel data into a serial bit stream and vice versa. The encoding process adds about 20 percent overhead to the data stream. PCI Express 1.0 has a signaling rate of 2.5 Gb/s per direction per lane, resulting in an effective maximum bandwidth of 250 MB/s per direction per lane after accounting for serial encoding overhead (Figure 9). Therefore, a x4 link — which has 4 lanes, each with a send and receive pair — has an effective bandwidth of 2 GB/s and a x8 link has an effective bandwidth of 4 GB/s.

Figure 9. PCI Express has an effective bandwidth of 250 MB/s after accounting for the overhead of serializing/deserializing encoding.



For additional information about PCI Express technology, see the technology brief titled "HP local I/O strategy for ProLiant servers."[11]

[11] Available on the HP ISS technology website at http://h18004.www1.hp.com/products/servers/technology/whitepapers/

Smart Array 6i controller

The ProLiant DL580 G3 includes the embedded Smart Array 6i controller. This integrated Smart Array controller is the first to support the Ultra320 SCSI interface which provides data transfer rates of up to 320 MB/s per channel. Because of the use of a common low-voltage differential signaling protocol, the Smart Array 6i supports Ultra2, Ultra3, and Ultra320 SCSI drives. Certain models of the ProLiant DL580 G3 ship standard with the 128 MB battery-backed write cache enabler (BBWC).

The Smart Array 6i controller is the next generation of the integrated Smart Array 5i Plus controller, using the faster PCI bus technology (64-bit, 133 MHz) and using twice the amount of memory in the BBWC.

The Smart Array 6i controller provides true hardware RAID protection, offering a cost-effective alternative to customers using software RAID or that do not use RAID at all. The Smart Array 6i controller supports RAID levels 0, 1, 0+1, and 5. Furthermore, HP designs the family of Smart Array controllers to be easy to use and easy to upgrade as business needs increase. If administrators need to expand their storage capabilities beyond the limits of the internal storage drives, they can upgrade simply from the 6i to an Ultra3 or Ultra320 Smart Array Controller. Refer to the best practice paper, "Enhancing performance of HP ProLiant servers containing Smart Array 6i controllers," for information about optimizing performance.[12]

For additional information about the Smart Array 6i controller, see the HP website.[13]

Mechanical design for serviceability

The ProLiant ML570 G3 and DL580 G3 servers follow the well-established tradition of excellent mechanical design in ProLiant servers. The designs improve thermal efficiencies, improve the ability to withstand mechanical shocks, and provide easy-to-service components. Also, because these servers were completely redesigned, HP was able to design many parts in common between the two servers to simplify spare parts requirements.

Common components

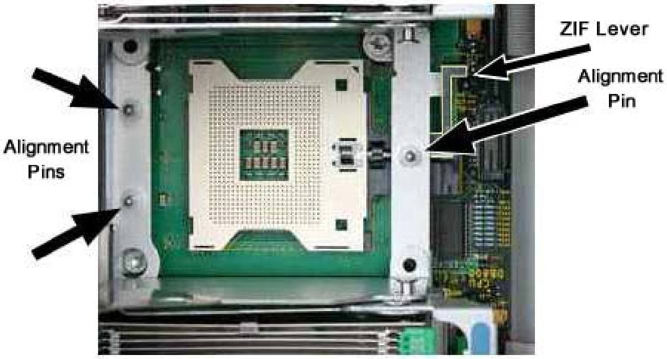
Both servers share the same processor heatsink design, hot-plug redundant fans, hot plug redundant power supply design, universal rail design, and the same cable management arms. Using the same parts allows the customer to keep fewer spare parts in stock and greatly simplifies ordering and storage of replacement parts.

[12] Available on the HP website at http://h20000.www2.hp.com/bc/docs/support/SupportManual/c00366606/c00366606.pdf

[13] See the ProLiant storage array controller webpage at http://h18004.www1.hp.com/products/servers/proliantstorage/arraycontrollers/index.html

The processor heat sink design includes three guiding pins and also a special locking lever mechanism to reduce the likelihood of damaging the processor when inserting the heat sink (Figure 10). The lever, which is extended out for easy access, has to be in the proper position before it can lock the heat sink down onto the processor.

Figure 10. Processor heat-sink design.

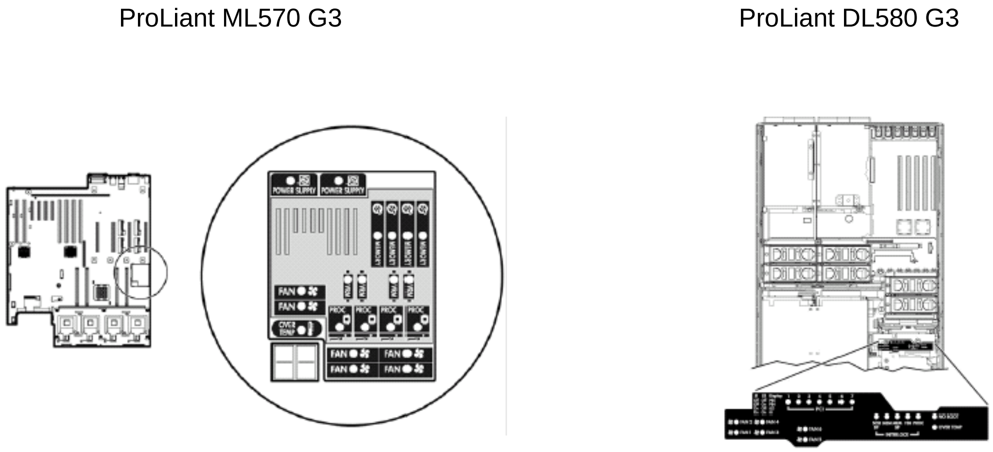


The ProLiant ML570 and the DL580 G3 use a single redundant, hot-plug fan form factor throughout the servers. Both servers support six hot-plug fans. The use of a single fan form-factor is a huge simplification compared to the G2 models, which had a total of five different fans between the two servers.

Both servers ship with tool-free “Snap In” sliding rails and a cable management system for simple deployment in HP and third-party racks and for in-rack server access. The rail system is ambidextrous so that the same rail design can be used on either side of the rack with minor adjustments to the rail system. The rail system provides tool-free support for square-hole, round-hole, and threaded-hole racks for adjusting to rack depths of 23 $^3/4$ to 35 inches — accommodating a much broader range of rack designs than previous rail systems. The cable management arms are also ambidextrous and independent of the hole design in the racks. The ambidextrous cable management arms allow customers to choose whether to put all the cables down one side of the rack or to switch sides from one server to the next.

Both servers incorporate a QuickFind diagnostic display inside the server chassis to provide specific trouble-shooting information (Figure 11). The display includes LEDs for all major subsystems of the server: PCI and PCI-X, memory, processors, redundant fans, overtemp indicator, etc. During normal operations, all of the LEDs are off. An amber LED gives instant visual indication of a fault condition.

Figure 11. Diagnostic displays for the ProLiant ML570 G3 (on the left side) and the ProLiant DL580 G3 (on the right side).



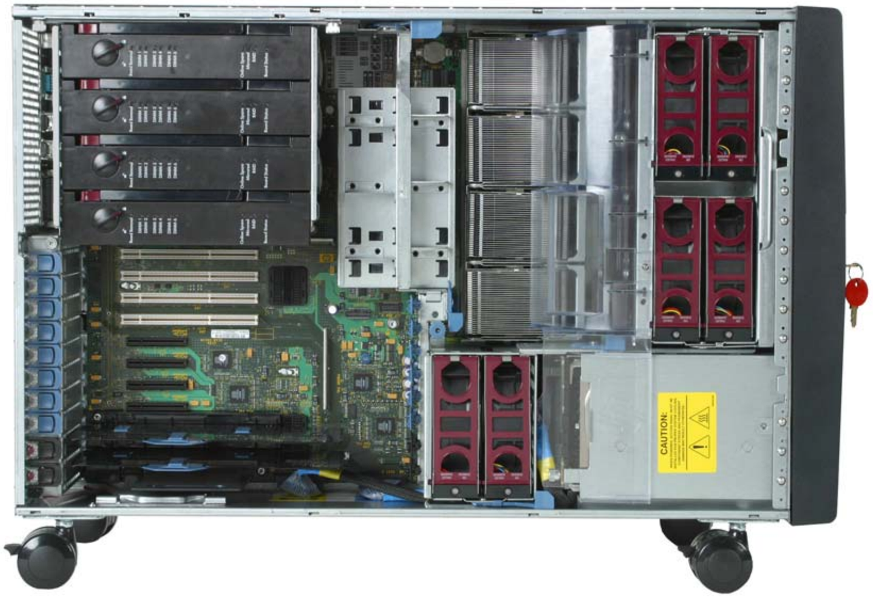
ProLiant ML570 G3

ProLiant DL580 G3

ProLiant ML570 G3

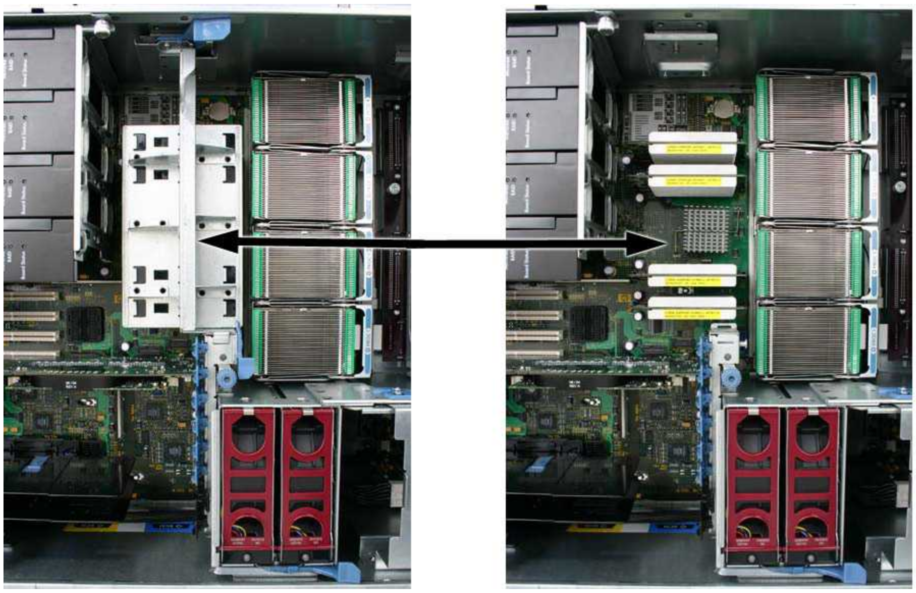
The ProLiant ML570 G3 is a more compact (6U) and lightweight design than the G2 model. It is almost a cable-less design, having half the number of cables that the G2 model had. The chassis is two levels, with most of the components on the top level (Figure 12). These components plug into the system board for easy access.

Figure 12. Internals for the ProLiant ML570 G3.



Administrators will find it easy to service the ProLiant ML570 G3. For example, the use of an innovative, split center wall reduces the number of steps required when adding a processor to the server. The center wall is needed for stability in the server to prevent flexing, especially during shipment. In the G3 model, the administrator can either remove the center wall (with the fans installed) as one piece or just remove the half of the center wall that is retaining the VRMs to update the processor and VRM (Figure 13).

Figure 13. A portion of the center wall (image shown on the left) can be removed (image on the right) to insert additional VRMs for the processors.

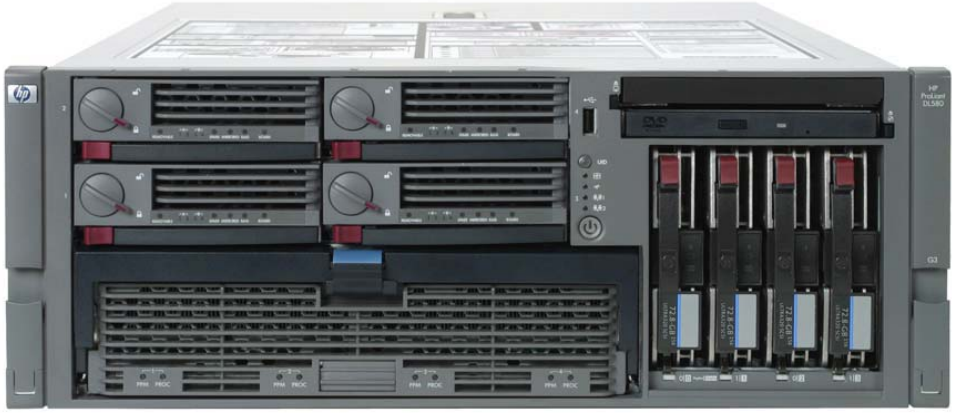


In the design of the ProLiant ML570 G3, care has been taken to remove the possibility of mishaps when an administrator is required to upgrade or service the server. For instance, the memory modules use an intuitive locking mechanism to prevent a user from inserting or removing a memory board when it is in the 'hot' or active state. When the administrator needs to add a processor, the heat sink on the processor has extensions that keep the processor pins from hitting any surfaces and being inadvertently bent. The power backplane and fan backplane are designed with components only on the top sides of the boards, to reduce any chance of failures due to components being knocked off or damaged on the bottom side of the boards. Similarly, for boards that do have components on both sides, subpans have been added to reduce the risk of damaging bottom-side components.

ProLiant DL580 G3

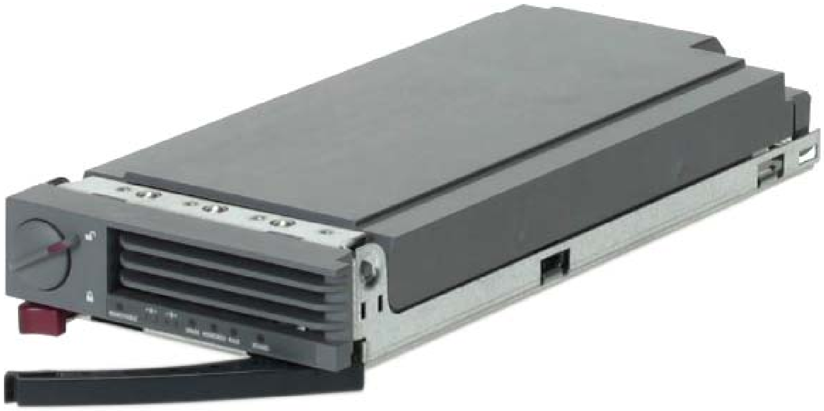
The ProLiant DL580 G3 provides a compact design for a powerful 4-way server. Like the ProLiant ML570 G3, the DL580 G3 uses plug-in components to eliminate many of the cables. One of the most important aspects of the design is that the all the memory modules, processors, and disk drives are easily accessible from the front of the unit (Figure 14).

Figure 14. Memory, processors, and disk drives are easily accessible on the ProLiant DL580 G3 server.



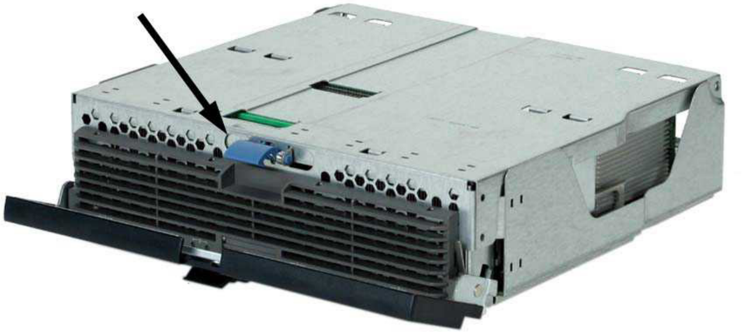
The memory modules use the same intuitive locking mechanism as on the ProLiant ML570 G3, to prevent a user from inserting or removing a memory board when it is in the 'hot' or active state (Figure 15).

Figure 15. Memory modules have a simple locking icon that shows whether the memory board can be removed.



The four processors are all housed in a processor cage that is accessible from the front, eliminating the need for the server to be removed from the rack to replace or install a processor. The processor cage uses a unique "double-locking" latch mechanism – customers use the blue touchpoint latch to remove the processor cage from the server chassis, then use this same blue touchpoint latch to open the top sheet metal cover to the processor cage (Figure 16).

Figure 16. Dual latch used for the processor cage.



Conclusion

The ProLiant ML570 G3 and the ProLiant DL580 G3 servers are the first enterprise class, 4-way servers to offer high performance combined with the increased reliability of Hot Plug RAID memory. The server designs share a common processor and memory architecture that optimizes performance through the use of the latest Intel 64-bit Xeon processor MP, fast DDR-2 400 memory, and multiple layers of memory interleaving to reduce latencies in the memory subsystem. The servers are capable of supporting a large memory footprint to complement the 64-bit capabilities of the Intel 64-bit Xeon processor MP. Advanced Memory Protection techniques protect against correctable memory errors (advanced ECC), reduce the risk of uncorrectable errors (Online Spare memory), and protect against uncorrectable errors (Hot Plug Mirrored Memory and Hot Plug RAID). The server chassis have been redesigned to share as many components as possible, reduce weight and complexity, and simplify serviceability for both the rack-optimized ProLiant DL580 G3 and the expansion-optimized ProLiant ML570 G3.

Appendix A. Engineering prefixes

Table A1

	Abbreviation	**Exponential form**	**Number of bytes**	**Relationship to next lowest prefix**
Gigabyte	[G/GB]	2^{30} bytes	1,073,741,824 bytes	1024 Megabytes
Terabyte	[T/TB]	2^{40} bytes	1,099,511,627,776 bytes	1024 Gigabytes
Petabyte	[P/PB]	2^{50} bytes	1,125,899,906,842,624 bytes	1024 Terabytes
Exabyte	[E/EB]	2^{60} bytes	1,152,921,504,606,846,976 bytes	1024 Petabytes

For more information

Resource description	**Web address**
HP Website **ProLiant server information** **ProLiant benchmarks** **Industry-Standard Server Technology Papers**	www.hp.com/go/proliant http://h18004.www1.hp.com/products/servers/benchmarks/index.html www.hp.com/servers/technology
Intel website	www.intel.com
JEDEC website **Memory standards and soft error information**	www.jedec.org

Call to action

To help us better understand and meet your needs for ISS technology information, please send comments about this paper to: TechCom@HP.com.

TC050601TB, 06/2005